日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　４月２１日

出願番号
Application Number: 特願２００３－１１５１８５

[ST. 10/C]: [ＪＰ２００３－１１５１８５]

出願人
Applicant(s): 株式会社日立製作所

２００３年　９月３０日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０８００９１

【書類名】　特許願
【整理番号】　K03008231A
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内
【氏名】　原　憲宏
【発明者】
【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内
【氏名】　宮崎　光夫
【発明者】
【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内
【氏名】　河村　信男
【発明者】
【住所又は居所】　神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内
【氏名】　八高　克志
【発明者】
【住所又は居所】　神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　ビジネスソリューション事業部内
【氏名】　菅　将孝
【特許出願人】
【識別番号】　000005108
【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100075096

【弁理士】

【氏名又は名称】　作田　康夫

【手数料の表示】

【予納台帳番号】　013088

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】

高可用性を提供するデータベース処理方法

【特許請求の範囲】

【請求項１】

データベースを複数の分割データベースに分割し、そのおのおのに対してＤＢサーバを関連付けてデータ処理を行うデータベース管理システムにおいて、

あらかじめダウン状態の他のＤＢサーバの処理を代行する関係を示す情報を登録しておくステップと、

あるＤＢサーバに処理を要求する際に、そのＤＢサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更するステップと、

その要求に代行するための指示を付加するステップと、

その要求を受け取った際に、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの替わりにデータ処理を行うステップと

を有することを特徴とするデータベース処理方法。

【請求項２】

請求項１に記載のデータベース処理方法において、

ダウンしているサーバの替わりにデータ処理を行う際に、実行環境をダウンしているサーバの実行環境に切り換えること

を特徴とするデータベース処理方法。

【請求項３】

請求項１に記載のデータベース処理方法において、

ダウンしているサーバの替わりにデータ処理を行う際に、本来のＤＢサーバと関連付けられたＤＢ格納領域、テーブルあるいはインデクスをアクセスするためのデータベースバッファを利用すること

を特徴とするデータベース処理方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明はデータ処理技術に関し、系切り替え機能を有するデータベース管理システムに適用して有効な技術に関するものである。

【０００２】

【従来の技術】

サービスの停止が重大なビジネスチャンスの喪失に直結するネットビジネスの世界では、２４時間３６５日稼動し続ける堅牢性を備えたシステムが求められる。その中でも特に、障害発生時の影響の局所化と迅速なシステム回復が重要である。従来より、データベース（ＤＢ）システムでは、障害発生時の迅速なシステム回復のために、サービス実行用の実行系マシンとは別に待機系マシンと用意し、障害発生時に待機系マシンにサービスの実行を切り替える「系切り替え」という技術が用いられてきた。

【０００３】

例えば、文献(Jim Gray and Andreas Reuter, Morgan Kaufmann Publishers, 1993)はＨＡ(High Availability)システム構成によるＤＢ障害対策としてホットスタンバイ無停止運用について開示している。

【０００４】

一方、文献Parallel Database Systems :The Future of High Performance Database Systems(David DeWitt and Jim Gray, COMMUNICATIONS OF THE ACM, Vol.35, No.6, 1992, P.85-P.98)には、データベースの処理負荷を複数のプロセッサに分散させ並列に実行するアーキテクチャが開示されている。上記従来技術に記載のShared everything, Shared disk（共用型）アーキテクチャでは、ＤＢ処理を実行する全ての計算機が全てのデータをアクセスすることが可能であり、Shared nothing（非共用型）アーキテクチャでは、自計算機に接続されたディスクに格納されたデータのみアクセス可能である。

【０００５】

Shared nothing（非共用型）アーキテクチャは、共用型アーキテクチャにくらべ、ＤＢ処理を実行する構成単位の間での共有リソースが少なく、スケーラビリティの点で非常に優れている。Shared nothing（非共用型）アーキテクチャにお

いても、高可用性を提供するために系切り換えの技術を用いることが多い。

【０００６】

【非特許文献１】

Jim Gray and Andreas Reuter, Morgan Kaufmann Publishers, 1993

【非特許文献２】

Parallel Database Systems :The Future of High Performance Database Systems(David DeWitt and Jim Gray, COMMUNICATIONS OF THE ACM, Vol.35, No.6, 1992, P.85-P.98)

【０００７】

【発明が解決しようとする課題】

しかし、系切り替えでは、実行系マシンとは別に待機系マシンとを用意しなければならず、通常サービス実行時には待機系マシンは遊んでいる状態である。また、系切り替えでも「相互待機」という形態で、待機系マシンにも通常のサービス実行を割り当てることもできるが、切り替え時のシステム回復の高速化を図るために、待機系でのシステム立上げをあらかじめ途中まで行っておく(「ウォームスタンバイ」、「ホットスタンバイ」)ことが多く、待機系システムのリソース(プロセスやメモリなど)を余分に用意しなければならない。

【０００８】

上記のような通常時未稼動状態である待機専用のリソースを必要とするシステムにおいては通常時にリソースを有効に利用できておらず、システム構築・運用におけるTCO(Total Cost of Ownership)削減の観点で問題である。

【０００９】

本発明の目的は上記問題を解決し、前述のような通常時未稼動状態である待機専用のリソースを必要とすることなく、かつ障害発生時に切り替え時間が短い、ＤＢシステム系切り替え制御方法を提供することである。特に、Shared nothingアーキテクチャを用いた並列データベース管理システムにおいて高可用性を提供するデータベース処理方法を提供することを目的とする。

【００１０】

【課題を解決するための手段】

前記課題を以下の手段により解決する。

【００１１】

Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて、あらかじめダウン状態の他のＤＢサーバの処理を代行する関係を示す情報を登録しておき、ユーザからの問合せを受け付け、あるＤＢサーバに処理を要求する際に、そのＤＢサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更する。そして、その要求に代行するための指示を付加する。上記要求を受け取ったＤＢサーバは、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの替わりにデータ処理を行う。

【００１２】

前記ダウンしているサーバの替わりにデータ処理を行う際には、替わりを務めるサーバは、実行環境を前記ダウンしているサーバの実行に切り換える。ここで、すでに以前替わりを務めた状態である場合には、実行環境の切り換えは行わない。また、データ処理を行う際のデータベースへのアクセスに使用するデータベースバッファは、本来のＤＢサーバに関連付けられたＤＢ格納領域へのアクセスに使用しようとしているデータベースバッファを共用する。上記ダウン状態の他のＤＢサーバの処理を代行する関係を示す情報は、データベース管理システムが内部で自動的に生成してもよい。さらに、代行の優先順位を付加することにより、あるＤＢサーバがダウンした場合の代行サーバを複数指定することができる。

【００１３】

【発明の実施の形態】

Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて予備系専用のリソースを用意することなく、障害発生時にＤＢアクセス処理を即座に再開することが可能な一実施形態のデータベース処理システムについて説明する。

【００１４】

まず、本発明の概念を図１を用いて簡単に説明する。

【００１５】

本実施例のデータベース管理システムは、処理要求受付サーバ（ＦＥＳ：Ｆｒｏｎｔ　Ｅｎｄ　Ｓｅｒｖｅｒ）１０およびＤＢアクセスサーバ（ＢＥＳ：Ｂａｃｋ　Ｅｎｄ　Ｓｅｒｖｅｒ）２０を具備する。

【００１６】

処理要求受付サーバ（ＦＥＳ）１０は、ユーザからの問合せ５０を受け付け解析し、ＤＢアクセス要求を生成し、ＤＢアクセスサーバへＤＢアクセス要求を行う。そしてＤＢアクセスの結果を必要に応じてユーザに返す。ＤＢアクセスサーバ（ＢＥＳ）２０は、処理要求受付サーバ１０からのＤＢアクセス要求を受け取り、要求にしたがってＤＢ格納領域上のデータを操作し、必要に応じて結果を処理要求受付サーバに返す。ＦＥＳ１０およびＢＥＳ２０とも１プロセスもしくは複数のプロセスによって実現される。

【００１７】

本実施例のデータベース管理システムのアーキテクチャは、Shared nothing(非共用型)アーキテクチャであり、本システムが管理するデータベース(例えばテーブル、インデクス)は、さまざまな手法により複数の分割テーブルおよび分割インデクスに分割され、複数のＤＢ格納領域に分散格納される。あるＤＢ格納領域は決まったＤＢアクセスサーバに対応付けられており、ＤＢアクセスサーバは、そのＤＢアクセスサーバに対応付けられたＤＢ格納領域内のデータ(例えばテーブルデータ、インデクスデータ)のみをアクセスする。

【００１８】

図１の例では、通常、ＢＥＳ１は、ＤＢ格納領域１へのアクセス要求のみを、ＢＥＳ２は、ＤＢ格納領域２へのアクセス要求のみを処理する。ＢＥＳ１およびＢＥＳ２は同じＤＢ格納領域をアクセスすることはない。

【００１９】

通常時、ＢＥＳ１およびＢＥＳ２とも稼動状態であり、すべてのリソース（ＤＢアクセスサーバを実現するプロセス、メモリ等）が有効に活用されている。

【００２０】

ここで例えば情報処理装置に電源等の障害が発生し、ＢＥＳ１がダウンした場合、ほかの稼動中のサーバ(本例ではＢＥＳ２)に処理を引き継いでサービスを続

行する。すなわち、ＢＥＳ２がＢＥＳ１に対するＤＢアクセス要求を代行する。

【００２１】

具体的には、ユーザ(すなわちアプリケーションプログラム)からの問合せ要求５０を受け取ったＦＥＳ（１０）は、問合せ要求を解析し、アクセスするデータが存在するＤＢ格納領域を決定する。決定されたＤＢ格納領域へのアクセスを担当するＤＢアクセスサーバがダウン中である場合、障害時の代行情報を基に、処理を代行してくれる代行サーバを決定する（１３）。本例では、ＢＥＳ１の代行サーバはＢＥＳ２であることから、ＢＥＳ２を代行先として決定する。そして、ＤＢアクセス要求４０に代行指示を付加し（１４）、ＤＢアクセスサーバＢＥＳ２に対してそのＤＢアクセス要求を送信する（１５）。代行指示の実装形態は、フラグでもＢＥＳ１を特定できる識別子でもかまわない。本例ではフラグとして説明する。

【００２２】

ＦＥＳ（1.0）からのＤＢアクセス要求４０を受け取ったＤＢアクセスサーバＢＥＳ２は、ＤＢアクセス要求がＢＥＳ２が担当しているＤＢ格納領域２への要求であるか、別のＤＢ格納領域へのアクセス要求であるかを判定する（２０）。判定には、ＤＢアクセス要求に付加された代行指示を用いる。代行指示のフラグがＯＮの場合、障害時の代行情報を基に本来ＢＥＳ１が担当しているＤＢ格納領域１へのアクセス要求であることを認識する。そして、ＤＢアクセスサーバＢＥＳ１の実行環境に切り換える（２４）。例えば、ＤＢアクセスサーバＢＥＳ２を実現しているプロセスの環境変数やメモリ情報をＢＥＳ１のものに変更する。

【００２３】

切り換えた実行環境の下、ＤＢアクセス処理を実行し（２５）、ＤＢ格納領域１をアクセスし要求されたデータ操作を行う。本例では、データベース６０内のＤＢ格納領域１（６１）に格納されたテーブルデータ（６２）の”12”を実行結果として、ＦＥＳに送信する。送信されたテーブルデータは問合せ結果としてＦＥＳ（１０）によってユーザに返却される。

【００２４】

ここでの障害時の代行情報は、ＤＢＭＳの管理者などのユーザによりあらかじ

めＤＢＭＳに登録されているものとするが、ＤＢＭＳが内部に自動生成することにより管理者の負担を低減することができる。

【００２５】

障害が発生した際、ダウン中のＤＢアクセスサーバＢＥＳ１の替わりにＤＢ格納領域１へのアクセス要求を代行して処理してくれるＤＢアクセスサーバＢＥＳ２をここでは、「代行サーバ」と呼ぶ。代行サーバとして動作する際に、あらたなプロセスが生成されるわけではなく、もともとのＢＥＳのプロセスのまま、代行サーバとして処理を行うため、無駄なプロセス生成もない。

【００２６】

以上のように、別のサーバを代行サーバとして登録もしくは決定しておき、障害発生時には、ＦＥＳでＢＥＳ未稼動状態を検知し、代行指示を用いて稼動中のサーバに処理の代行をさせることにより、予備系専用のリソースなしで、障害発生時に即座にＤＢアクセス処理を再開することが可能である。

【００２７】

ここでは、ＦＥＳとＢＥＳとを別々の情報処理装置上に配置したが、同一の情報処理装置上に配置することによりハードウェア資源の有効活用が可能である。また、本実施例で示したＦＥＳの機能およびＢＥＳの機能を１つのＤＢサーバとして実装することにより、データベース管理システムの管理者は、ＦＥＳとＢＥＳとを意識して管理する必要がなくなり、管理コストを低減することができる。

【００２８】

次に図２に本実施形態のデータベース管理システムの概略構成を示す。

【００２９】

ユーザが作成したアプリケーションプログラム６と、問い合わせやリソース管理などのデータベースシステム全体の管理を行うデータベース管理システム２がある。上記のデータベース管理システム２は、処理要求受付サーバ（ＦＥＳ）１０、ＤＢアクセスサーバ（ＢＥＳ）２０を具備する。また、データベース管理システム２はデータベースバッファ２３０を具備し、データベースアクセス対象となるデータを永続的にあるいは一時的に格納するデータベース３、そして障害時の代行情報３０を有する。

【００３０】

処理要求受付サーバ（ＦＥＳ）１０は、アプリケーションプログラム６から投入される問合せを受け付け解析し、ＤＢアクセス要求を生成し、ＤＢアクセスサーバへＤＢアクセス要求を行う。そしてＤＢアクセスの結果を必要に応じてアプリケーションプログラム６に返す。ＤＢアクセスサーバ（ＢＥＳ）２０は、処理要求受付サーバ１０からのＤＢアクセス要求を受け取り、要求にしたがって外部記憶装置上に記憶されるデータベース３をデータベースバッファ２３０を通じてアクセスする。先に図１において説明した代行サーバのＤＢアクセス処理では、もともとのＢＥＳが使用しているものと同じデータベースバッファを利用する。すなわち、もともとのＢＥＳと代行サーバ（ＢＥＳ）とではデータベースバッファを共用する。

【００３１】

上記データベース管理システム２はネットワークなどを介して他のシステムと接続されている。また、処理要求受け付けサーバ（ＦＥＳ）１０およびＤＢアクセスサーバ（ＢＥＳ）２０は必ずしも同一の情報処理装置上に配置される必要はない。それぞれ別々の情報処理装置上に配置され、ネットワーク等を介して１つのデータベース管理システムとして機能すればよい。また、１つのデータベース管理システムは複数のＦＥＳを配置することにより多量のユーザからの要求の負荷を分散させることができる。また、複数のＢＥＳを有することによりデータ処理の並列度が高まり、大規模なデータベースに対するデータ処理も高速に実現することができる。

【００３２】

上記処理要求受付サーバ１０は、問合せの構文解析・意味解析を行い、適切な処理手順を決定し、その処理手順に対応したコードを生成し、ＤＢアクセスサーバ２０に対しＤＢアクセス要求を行う処理要求制御部２１１を具備している。また、上記処理要求受付サーバ１０は、処理要求制御部２１１においてＤＢアクセス要求を行う際に、要求先ＤＢアクセスサーバの稼動状況を判定し、必要に応じてデータ処理要求先を変更し（２１３）、変更先に代行を支持する（２１４）代行制御部Ｆ（２１２）を具備している。

【００３３】

上記ＤＢアクセスサーバ２０は、処理要求受付サーバ１０から受け取ったＤＢアクセス要求（生成したコード）にしたがってデータベース３上のデータのアクセス制御等を行うデータ処理制御部２２１を具備している。また、上記ＤＢアクセスサーバ２０は、データ処理制御部２２１においてＤＢアクセス要求を受け取った際に、代行処理要求であるかを判定し、代行サーバとしての実行環境への切り替え（２２３）等の制御を行う代行処理制御部Ｂ（２２２）を具備している。また、代行処理制御部Ｂは、代行元のサーバに障害が発生した際に、関連するＤＢ格納領域ほかの回復を行い、実行中だった処理の更新結果を取り消すトランザクション回復を行う代行部回復制御機能を有する。

【００３４】

図３は本実施形態のコンピュータシステムのハードウエア構成の一例を示す図である。

この例のコンピュータシステムは、情報処理装置３０００、３１００および３２００を含む。

【００３５】

情報処理装置３０００は、ＣＰＵ３００２、主記憶装置３００１、通信制御装置３００３、Ｉ／Ｏ制御装置３００４及び端末３００６により構成される。主記憶装置３００１上には、アプリケーションプログラム３００８が置かれ、ＣＰＵ３００２を用いて稼動している。アプリケーションプログラム３００８がＤＢＭＳ２の処理要求受付サーバ１０にユーザ問合せ５０を行うと、情報処理装置３０００の通信制御装置３００３と情報処理装置３１００の通信制御装置３００３によって、ネットワーク３００７を経由して処理要求受付サーバ１０に問合せ要求が送られる。

【００３６】

情報処理装置３１００は、ＣＰＵ３００２、主記憶装置３００１、通信制御装置３００３、Ｉ／Ｏ制御装置３００４、磁気ディスク装置等の外部記憶装置３００５及び端末３００６により構成される。情報処理装置３１００の主記憶装置３００１上には、図２を用いて先に説明した処理要求受付サーバ１０を有するデー

タベース管理システム２が置かれ、ＣＰＵ３００２を用いて稼動している。外部記憶装置３００５上にはデータベース管理システム３が管理するデータベース３が格納される。また、データベース管理システム２を実現するプログラム３１００も外部記憶装置３００３上に格納される。処理要求受付サーバ１０は、Ｉ／Ｏ制御装置装置３００４により外部記憶装置３００５からデータの読み出し/書き込みを行い、通信制御装置３００３によりネットワーク３００７で接続された他の情報処理装置とデータの送受信を行う。

【００３７】

情報処理装置３２００は、ＣＰＵ３００２、主記憶装置３００１、通信制御装置３００３、Ｉ／Ｏ制御装置３００４、磁気ディスク装置等の外部記憶装置３００５及び端末３００６により構成される。情報処理装置３２００の主記憶装置３００１上には、図２を用いて先に説明したＤＢアクセスサーバ２０を有するデータベース管理システム２が置かれ、ＣＰＵ３００２を用いて稼動している。外部記憶装置３００５上にはデータベース管理システム３が管理するデータベース３が格納される。また、データベース管理システム２を実現するプログラム３１００も外部記憶装置３００３上に格納される。ＤＢアクセスサーバ２０は、Ｉ／Ｏ制御装置装置３００４により外部記憶装置３００５からデータの読み出し/書き込みを行い、通信制御装置３００３によりネットワーク３００７で接続された他の情報処理装置とデータの送受信を行う。

【００３８】

ここで、２つの情報処理装置３２００にまずそれぞれ関連付けられたデータベース３が格納された外部記憶装置３００５は、共用ディスクであり、他方の情報処理装置からのアクセスも可能である。データベース管理システム２の稼動および系切り替え操作を制御するクラスタウエアなどにより、上記共用ディスクのアクセス制御は行われる。

【００３９】

図４は本実施形態の処理要求制御部および代行制御部Ｆの処理手順の一部を示すフローチャートである。

【００４０】

まず、ステップ４０１において、ユーザからの問合せを受け取り、ステップ４０２において、ユーザからの問合せを解析して得られた情報から、問合せを実現するためのＤＢ格納領域へのアクセスを担当しているＤＢアクセスサーバを確定する。次にステップ４０３において、確定したＤＢアクセスサーバが稼動中であるかどうかを判定する。そのＤＢアクセスサーバが稼動中である場合、ステップ４０７に進み、確定したＤＢアクセスサーバに対して処理要求を送信する。

【００４１】

また、ステップ４０３において、確定したＤＢアクセスサーバがダウンしていて未稼動状態である場合、ステップ４０４に進み、当該ＤＢアクセスサーバに関連する障害時の代行情報を取得する。そして、取得した障害時の代行情報３０から代行サーバを代行先に決定する（ステップ４０５）。次に、処理要求へ代行指示を付加、すなわち代行指示のフラグをＯＮにする（ステップ４０６）。ステップ４０７に進み、代行先に決定したＤＢアクセスサーバに対して処理要求を送信する。

【００４２】

図５は本実施形態のデータ処理制御部および代行制御部Bの処理手順の一部を示すフローチャートである。

【００４３】

ＤＢアクセス処理は、図４のフローチャートで説明した処理要求の送信に引き続き行われる。

【００４４】

まず、ステップ５０１において、ＦＥＳから実行要求を受け取り、受け取った実行要求が代行指示かどうかの判定を行う（ステップ５０２）。代行指示のフラグがＯＮの場合、ステップ５０３に進み、自ＤＢアクセスサーバが代行サーバに指定されている障害時の代行情報を取得する（ステップ５０３）。そして、取得した障害時の代行情報３０から代行指示のあったサーバ名を参照し、そのサーバに関する実行環境へ切り換える（ステップ５０４）。切り換えた実行環境の下でＤＢアクセス処理を実行し、ダウン中のＤＢアクセスサーバに代わってそのＤＢ格納領域をアクセスしデータ操作を行う（ステップ５０５）。

【００４５】

また、ステップ５０２の判定において代行指示のフラグがＯＦＦの場合には、実行環境の切り換えは行わず、もともとのＤＢアクセスサーバとして自分の担当するＤＢ格納領域をアクセスし、データ操作を行う（ステップ５０５）。

【００４６】

ステップ５０４において、稼動中の実行環境のチェックを行い、処理要求を実行するためのＢＥＳと、現稼動中のＢＥＳとが同じ場合には、実行環境を切り替えないような制御を行うことも考えられる。また、ステップ５０５においてＤＢアクセス処理実行後に、もともとのＢＥＳの実行環境に戻すことも考えられる。それらは、処理要求を制御するバランサもしくはスケジューラによって最適な制御を行うことが可能である。

【００４７】

図６は本実施形態の代行情報の一例を示す図である。

図６（ａ）の例では、障害時の代行情報３０は、ＤＢアクセスサーバのサーバ名と、そのサーバがダウンした場合に、対応するＤＢ格納領域へのアクセスを替わりに行う代行サーバのサーバ名より構成される。代行情報６００は、ＤＢアクセスサーバＢＥＳ１が障害によりダウンした場合に、ＢＥＳ２が代行サーバとして処理を代行することを示している。

【００４８】

また、図７および図８は、図６の（ｂ）、（ｃ）、（ｄ）および（ｅ）にそれぞれ対応した代行サーバ構成を示す図である。

【００４９】

図６（ｂ）の代行情報の例は、図７（ａ）の代行サーバ構成に対応している。図６（ｂ）の例では、情報処理装置７０１上のＤＢアクセスサーバＢＥＳ３（７０５）と、情報処理装置７０２上のＤＢサクセスサーバＢＥＳ４（７０６）は、図６（ｃ）の代行情報６０１および６０２によって、それぞれ相互に代行サーバとして指示されている。すなわち相互代行の構成をとっている。具体的には、ＢＥＳ３がダウンした場合には、ＢＥＳ３の処理をＢＥＳ４が代行する（代行情報６０１）。ＢＥＳ４がダウンした場合には、ＢＥＳ４の処理をＢＥＳ３が代行す

る（代行情報６０２）。

【００５０】

さらに、情報処理装置７０３上のＢＥＳ５およびＢＥＳ６と、情報処理装置７０４上のＢＥＳ７およびＢＥＳ８も、２つの情報処理装置間でそれぞれのＢＥＳが相互代行構成をとっている。具体的には、ＢＥＳ５がダウンした場合には、ＢＥＳ７が処理を代行し、ＢＥＳ６がダウンした場合には、ＢＥＳ８が処理を代行する。また、ＢＥＳ７がダウンした場合には、ＢＥＳ５が処理を代行し、ＢＥＳ８がダウンした場合には、ＢＥＳ６が処理を代行する。

【００５１】

図７（ｂ）の表記において、例えば７０５の表記“（ＢＥＳ４＊）”の‘(’および‘)’は、そのＤＢサーバが“ＢＥＳ４”としては稼動していない、すなわちＤＢ格納領域ＲＤ４をアクセスするＤＢ処理は行っていない（ＢＥＳ３として稼動している）ことを示す。また、表記‘＊’は、ＤＢサーバ７０５が表記‘＊’を付加したＤＢサーバ（ＢＥＳ４）の代行を行うよう指定されていることを示す。

【００５２】

図７（ａ）におけるサーバの稼動状況は、情報処理装置７０１および７０３に障害が発生しＢＥＳ３、ＢＥＳ５およびＢＥＳ６がダウンした状態である。ＤＢ格納領域ＲＤ３へのアクセスを伴うデータ操作要求は、代行指示を付加されＤＢアクセスサーバ７０６に送信される。要求を受け取ったＤＢアクセスサーバ７０６のプロセスは、ＢＥＳ３として処理を代行しＲＤ３をアクセスする。

【００５３】

ＲＤ５へのアクセスを伴うデータ操作要求も同様に、代行指示を付加されＤＢアクセスサーバ７０７に送信される。要求を受け取ったＤＢアクセスサーバ７０７のプロセスは、ＢＥＳ５として処理を代行しＲＤ５をアクセスする。

【００５４】

図６（ｃ）の代行情報の例は、図７（ｂ）の代行サーバ構成に対応している。図６（ｃ）の例では、情報処理装置７１２上のＤＢアクセスサーバＢＥＳ９（７１５）、情報処理装置７１３上のＤＢサクセスサーバＢＥＳ１０（７１６）およ

びＤＢサクセスサーバＢＥＳ１１（７１７）は、図６（ｃ）の代行情報６０７、６０８および６０９によって、それぞれ片方向に代行サーバとして指示されている。すなわち片方向代行のリング構成をとっている。具体的には、ＢＥＳ９がダウンした場合には、ＢＥＳ９の処理をＢＥＳ１０が代行する（代行情報６０７）。ＢＥＳ１０がダウンした場合には、ＢＥＳ１０の処理をＢＥＳ１１が代行する（代行情報６０８）。そして、ＢＥＳ１１がダウンした場合には、ＢＥＳ１１の処理をＢＥＳ９が代行する（代行情報６０９）。

【００５５】

図７（ｂ）におけるサーバの稼動状況は、情報処理装置７１２に障害が発生しＢＥＳ９がダウンした状態である。ＤＢ格納領域ＲＤ９へのアクセスを伴うデータ操作要求は、代行指示を付加されＤＢアクセスサーバ７１６に送信される。要求を受け取ったＤＢアクセスサーバ７１６のプロセスは、ＢＥＳ９として処理を代行しＲＤ９をアクセスする。

【００５６】

図６（ｄ）の代行情報の例は、図８（ａ）の代行サーバ構成に対応している。図６（ｄ）の例では、情報処理装置８０１上のＤＢアクセスサーバＢＥＳ１２（８０４）、情報処理装置８０２上のＤＢサクセスサーバＢＥＳ１２（８０５）およびＤＢサクセスサーバＢＥＳ１４（８０６）は、図６（ｄ）の代行情報６１０、６１１および６１２によって、それぞれ代行サーバが指示されている。特に、代行情報６１１および６１２によって、ＢＥＳ１２はＢＥＳ１３およびＢＥＳ１４の２つのサーバの代行サーバとして指定されている。すなわち、Ｎ：１の代行構成になっている。

【００５７】

図８（ａ）におけるサーバの稼動状況は、情報処理装置８０３に障害が発生しＢＥＳ１４がダウンした状態である。ＤＢ格納領域ＲＤ１４へのアクセスを伴うデータ操作要求は、代行指示を付加されＤＢアクセスサーバ８０１に送信される。要求を受け取ったＤＢアクセスサーバ８０１のプロセスは、ＢＥＳ１４として処理を代行しＲＤ１４をアクセスする。

【００５８】

図６（ｅ）の代行情報の例は、図８（ｂ）の代行サーバ構成に対応している。図６（ｅ）の例では、情報処理装置８０８上のＤＢアクセスサーバＢＥＳ１２（８０４）は、図６（ｄ）の代行情報６１３、６１４および６１５によって、ＢＥＳ１３（８０５）、ＢＥＳ１４（８０６）およびＢＥＳ１５（８０７）を代行サーバとして複数指示している。ＢＥＳ１２がダウンしている場合、ＢＥＳ１２の処理を代行するサーバは、代行情報の代行優先順の値によって決定する。すなわちＮ段代行構成になっている。

【００５９】

まずは、代行情報６１３の指定により、ＢＥＳ１６が代行サーバの候補となる。ここでもしＢＥＳ１６もダウンしている場合は、代行優先順に従い代行情報６１４のＢＥＳ１７が代行サーバの候補となる。さらにＢＥＳ１７もダウンしている場合には同様にＢＥＳ１８が代行サーバとなる。

【００６０】

図８（ｂ）におけるサーバの稼動状況は、情報処理装置８０８に障害が発生しＢＥＳ１５がダウンした状態である。さらに情報処理装置８０９の障害によってＢＥＳ１６もダウンしているため、ＤＢ格納領域ＲＤ１５へのアクセスを伴うデータ操作要求は、代行情報６１４に従い代行指示を付加されＤＢアクセスサーバ８１４に送信される。要求を受け取ったＤＢアクセスサーバ８１４のプロセスは、ＢＥＳ１５として処理を代行しＲＤ１５をアクセスする。

【００６１】

以上図４および図５で示したフローチャートの処理は、図３で例として示したコンピュータシステムにおけるプログラムとして実行される。しかし、そのプログラムは図３の例の様にコンピュータシステムに物理的に直接接続される外部記憶装置に格納されるものと限定はしない。ハードディスク装置、フレキシブルディスク装置等のコンピュータで読み書きできる記憶媒体に格納することができる。また、ネットワークを介して図３のコンピュータシステムを構成する情報処理装置とは別の情報処理装置に接続される外部記憶装置に格納することもできる。

【００６２】

【発明の効果】

本発明によれば、Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて、通常時に稼動していない待機専用のリソースを有することなく、障害発生時にＤＢ処理サービスをすぐに再開することが可能なデータベース処理方法およびシステムを提供することができる。

【図面の簡単な説明】

【図１】

本発明の概念図である。

【図２】

本実施形態のデータベース処理システムの機能ブロックを示す図である。

【図３】

本実施形態のコンピュータシステムのハードウエア構成の一例を示す図である。

【図４】

本実施形態の処理要求制御部および代行制御部Ｆの処理手順の一部を示すフローチャートである。

【図５】

本実施形態のデータ処理制御部および代行制御部Ｂの処理手順の一部を示すフローチャートである。

【図６】

本実施形態の代行情報の一例を示す図である。

【図７】

本実施形態の代行サーバ構成の一例を示す図である。

【図８】

本実施形態の代行サーバ構成の一例を示す図である。

【符号の説明】

１０…処理要求受付サーバ

２０…ＤＢアクセスサーバ

３０…障害時の代行情報

４０…ＤＢアクセス要求

４１…代行指示

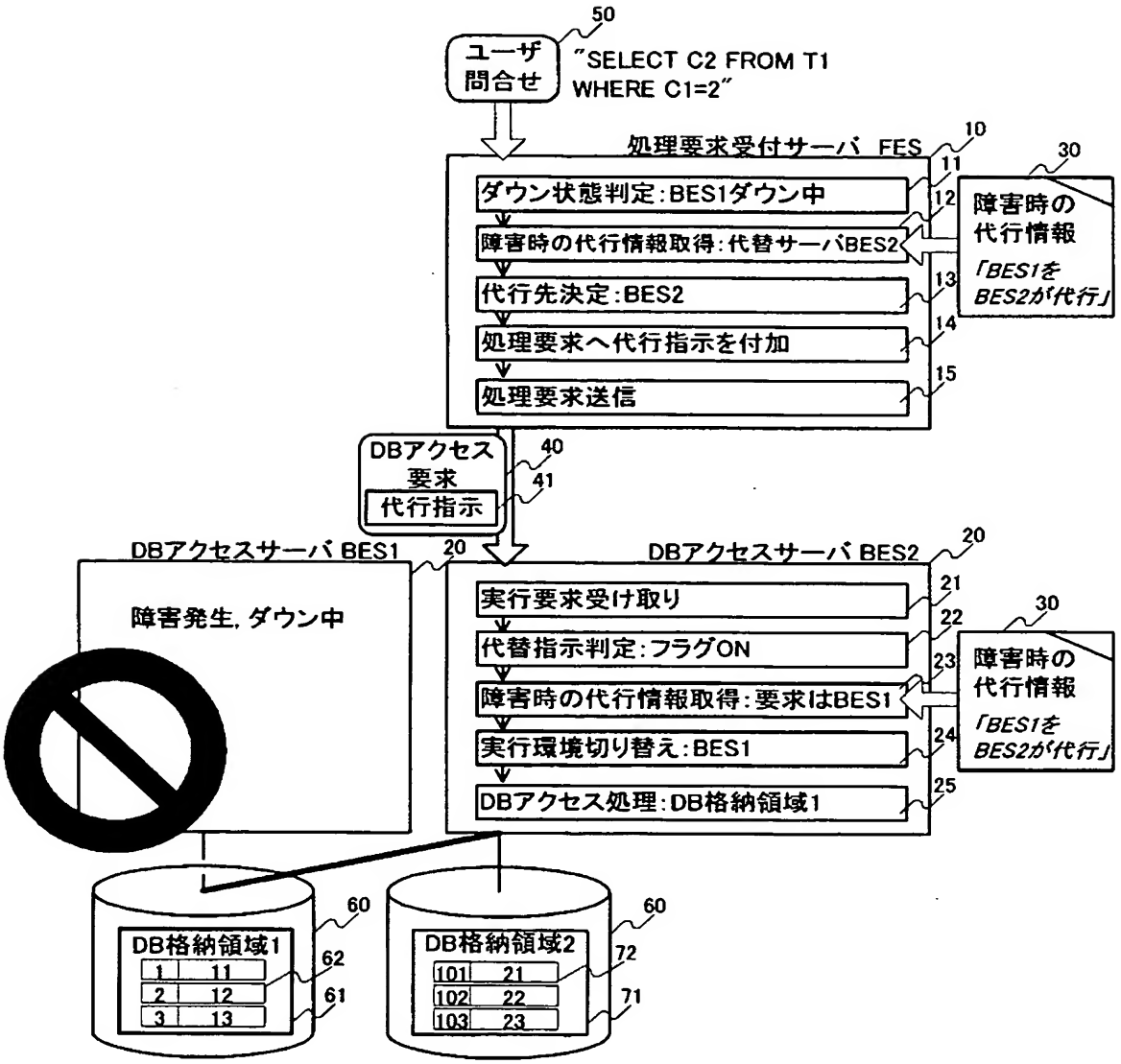
５０…ユーザ問合せ

６０…データベース。

【書類名】　図面

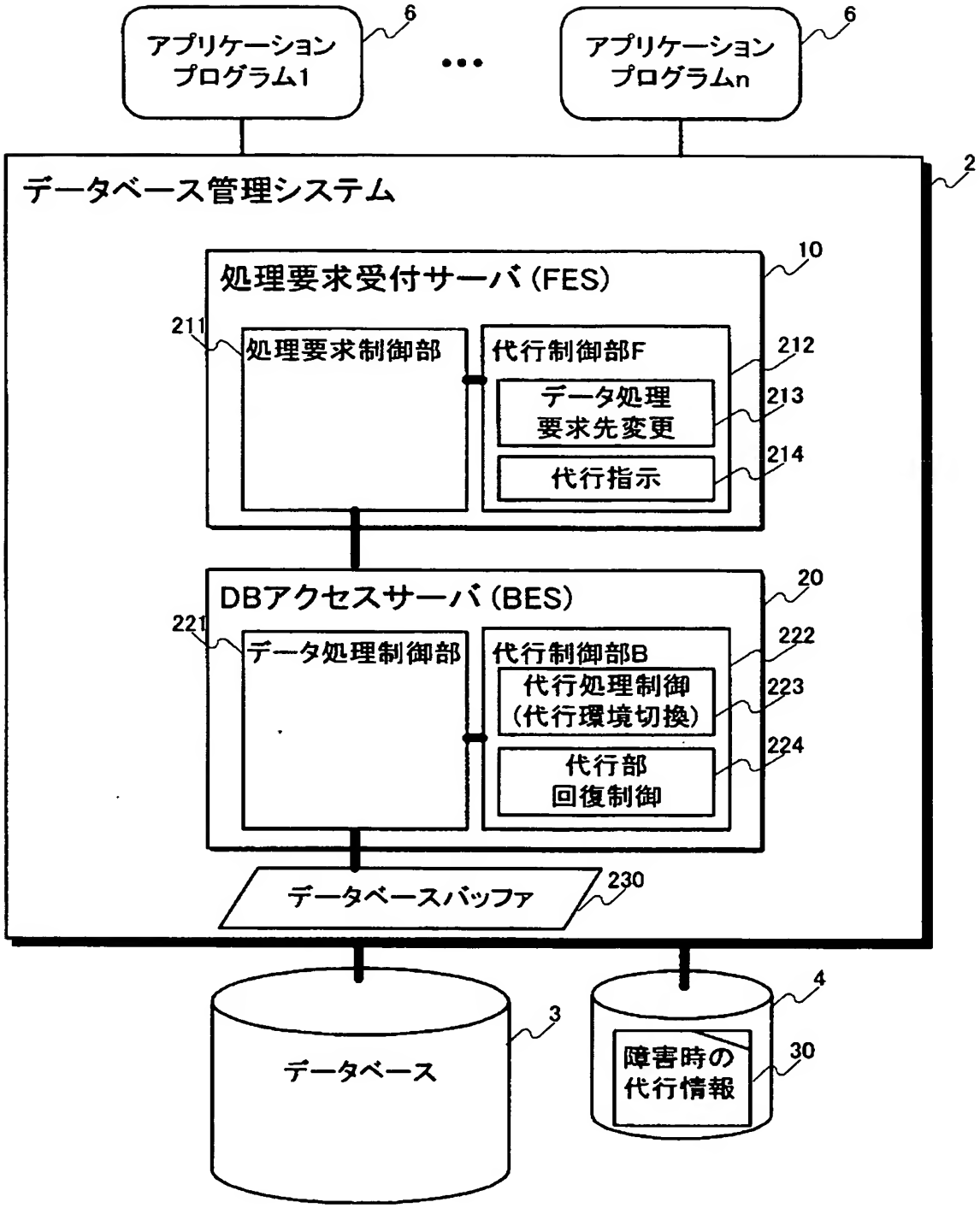
【図１】

【図1】


50
ユーザ
問合せ
"SELECT C2 FROM T1
WHERE C1=2"
処理要求受付サーバ FES
10
11
ダウン状態判定:BES1ダウン中
12
障害時の代行情報取得:代替サーバBES2
13
代行先決定:BES2
14
処理要求へ代行指示を付加
15
処理要求送信
30
障害時の
代行情報
「BES1を
BES2が代行」
DBアクセス
要求
代行指示
40
41
DBアクセスサーバ BES1
20
障害発生, ダウン中
DBアクセスサーバ BES2
20
21
実行要求受け取り
22
代替指示判定:フラグON
23
障害時の代行情報取得:要求はBES1
24
実行環境切り替え:BES1
25
DBアクセス処理:DB格納領域1
30
障害時の
代行情報
「BES1を
BES2が代行」
60
DB格納領域1
1 11
2 12
3 13
62
61
60
DB格納領域2
101 21
102 22
103 23
72
71

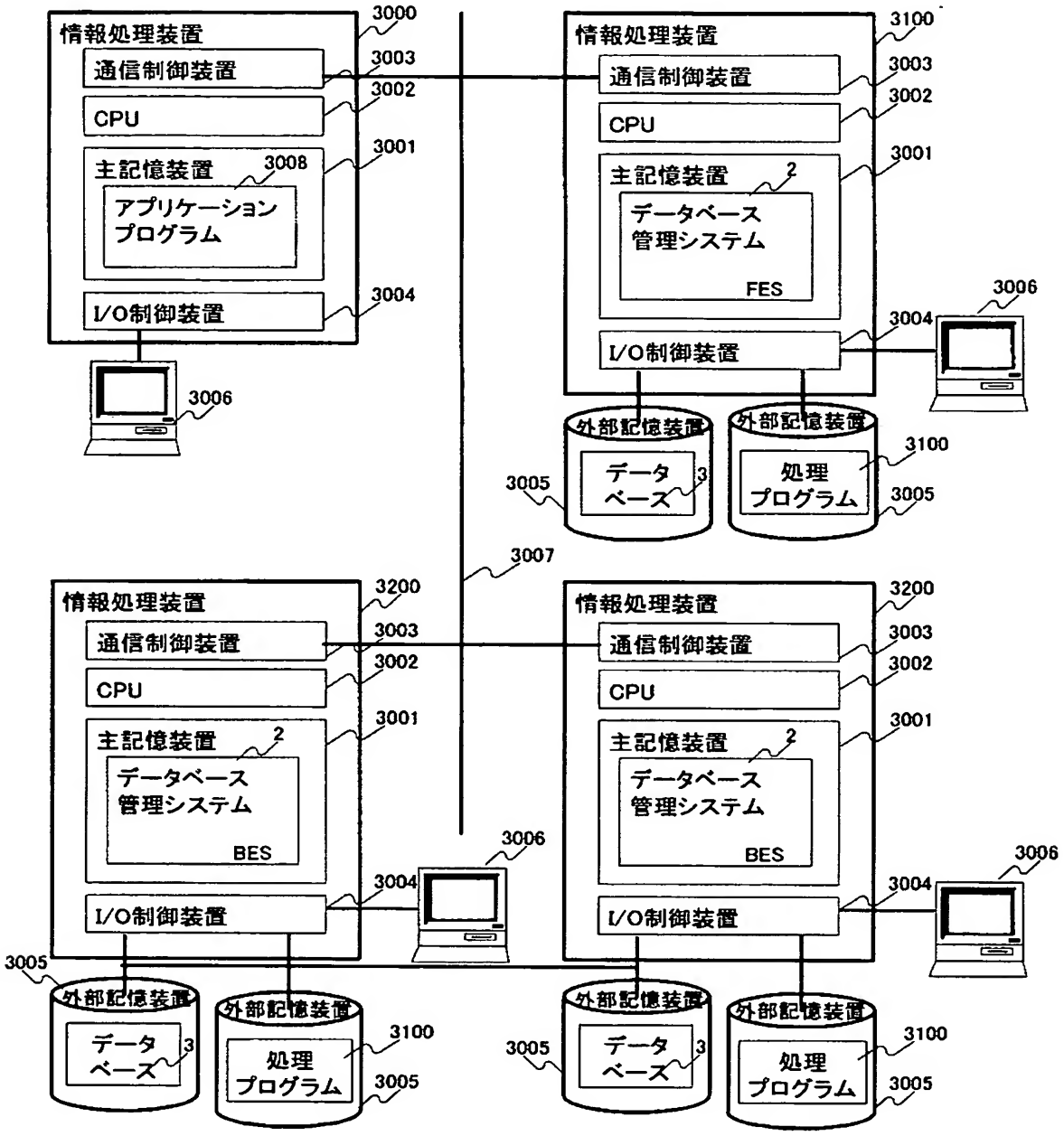
【図２】

【図2】


アプリケーション
プログラム1
6
アプリケーション
プログラムn
6
データベース管理システム
2
処理要求受付サーバ（FES）
10
211
処理要求制御部
代行制御部F
212
データ処理
要求先変更
213
代行指示
214
DBアクセスサーバ（BES）
20
221
データ処理制御部
代行制御部B
222
代行処理制御
（代行環境切換）
223
代行部
回復制御
224
データベースバッファ
230
データベース
3
障害時の
代行情報
4
30

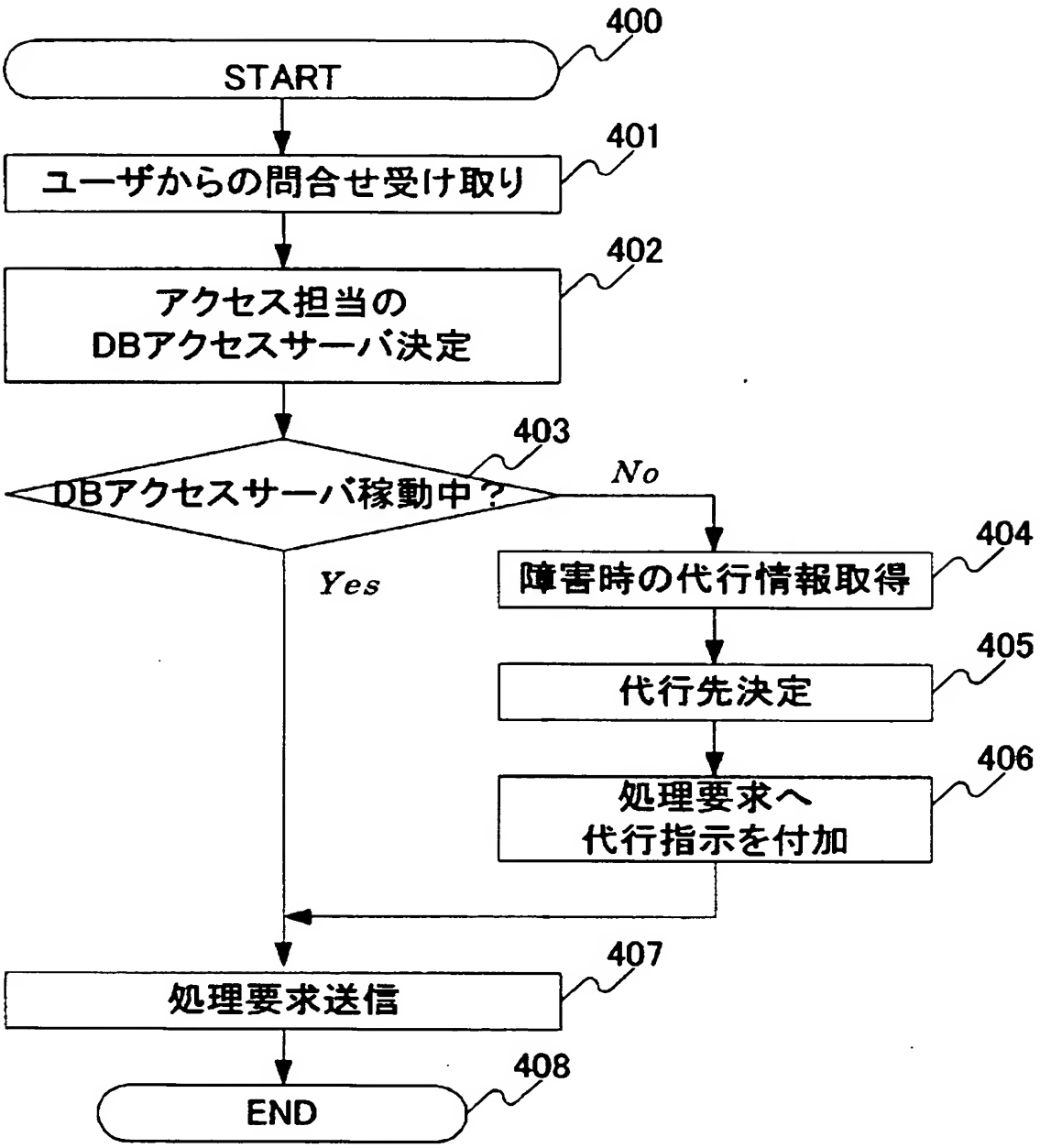
【図３】

【図3】


情報処理装置
3000
通信制御装置
3003
CPU
3002
主記憶装置
3008
アプリケーション
プログラム
3001
I/O制御装置
3004
3006
情報処理装置
3100
通信制御装置
3003
CPU
3002
主記憶装置
2
データベース
管理システム
FES
3001
I/O制御装置
3004
3006
外部記憶装置
データ
ベース
3
3005
外部記憶装置
処理
プログラム
3100
3005
3007
情報処理装置
3200
通信制御装置
3003
CPU
3002
主記憶装置
2
データベース
管理システム
BES
3001
I/O制御装置
3004
3006
情報処理装置
3200
通信制御装置
3003
CPU
3002
主記憶装置
2
データベース
管理システム
BES
3001
I/O制御装置
3004
3006
3005
外部記憶装置
データ
ベース
3
外部記憶装置
処理
プログラム
3100
3005
外部記憶装置
データ
ベース
3
3005
外部記憶装置
処理
プログラム
3100
3005

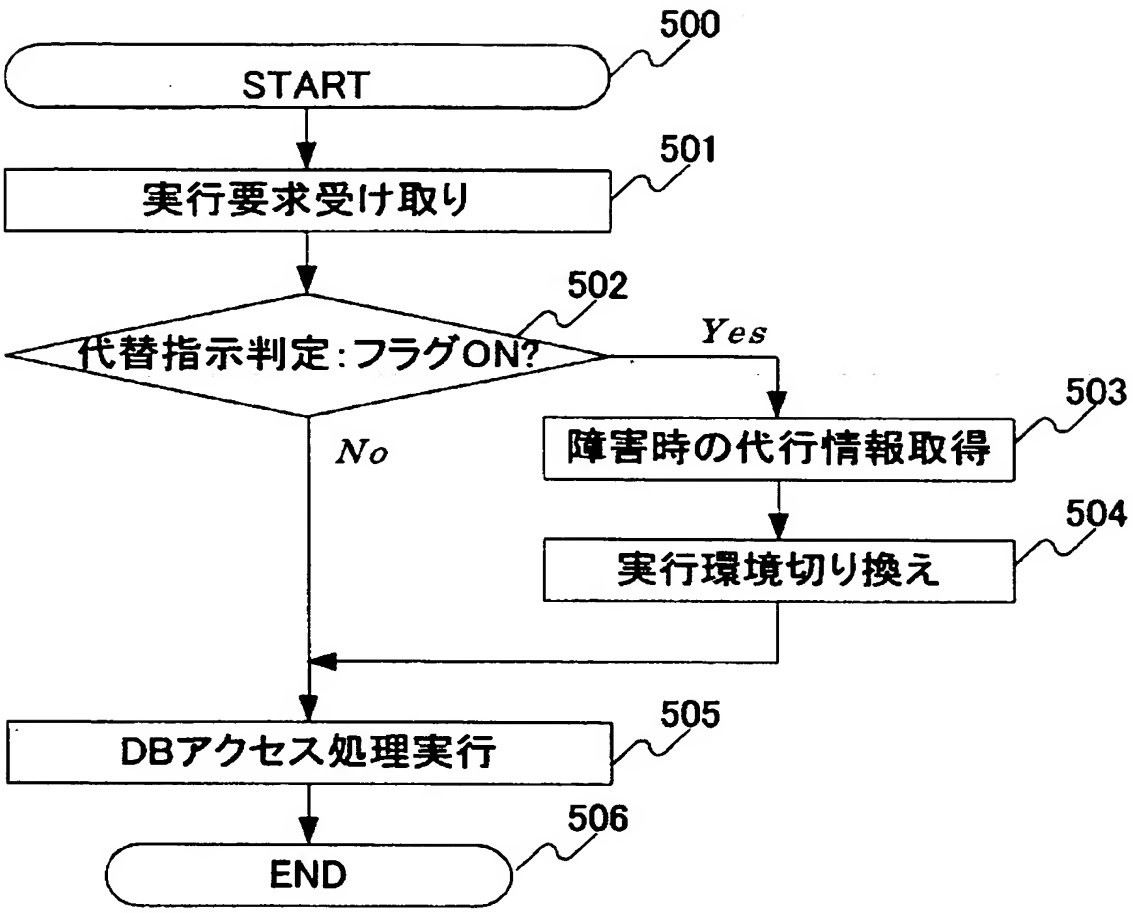
【図４】

【図4】


400
START
401
ユーザからの問合せ受け取り
402
アクセス担当の
DBアクセスサーバ決定
403
DBアクセスサーバ稼動中？
No
Yes
404
障害時の代行情報取得
405
代行先決定
406
処理要求へ
代行指示を付加
407
処理要求送信
408
END

【図５】

【図5】


500
START
501
実行要求受け取り
502
代替指示判定:フラグON?
Yes
No
503
障害時の代行情報取得
504
実行環境切り換え
505
DBアクセス処理実行
506
END

【図６】

【図6】

(a)

サーバ	代行サーバ	
BES1	BES2	600

(b)

サーバ	代行サーバ	
BES3	BES4	601
BES4	BES3	602
BES5	BES7	603
BES6	BES8	604
BES7	BES5	605
BES8	BES6	606

(c)

サーバ	代行サーバ	
BES9	BES10	607
BES10	BES11	608
BES11	BES9	609

(d)

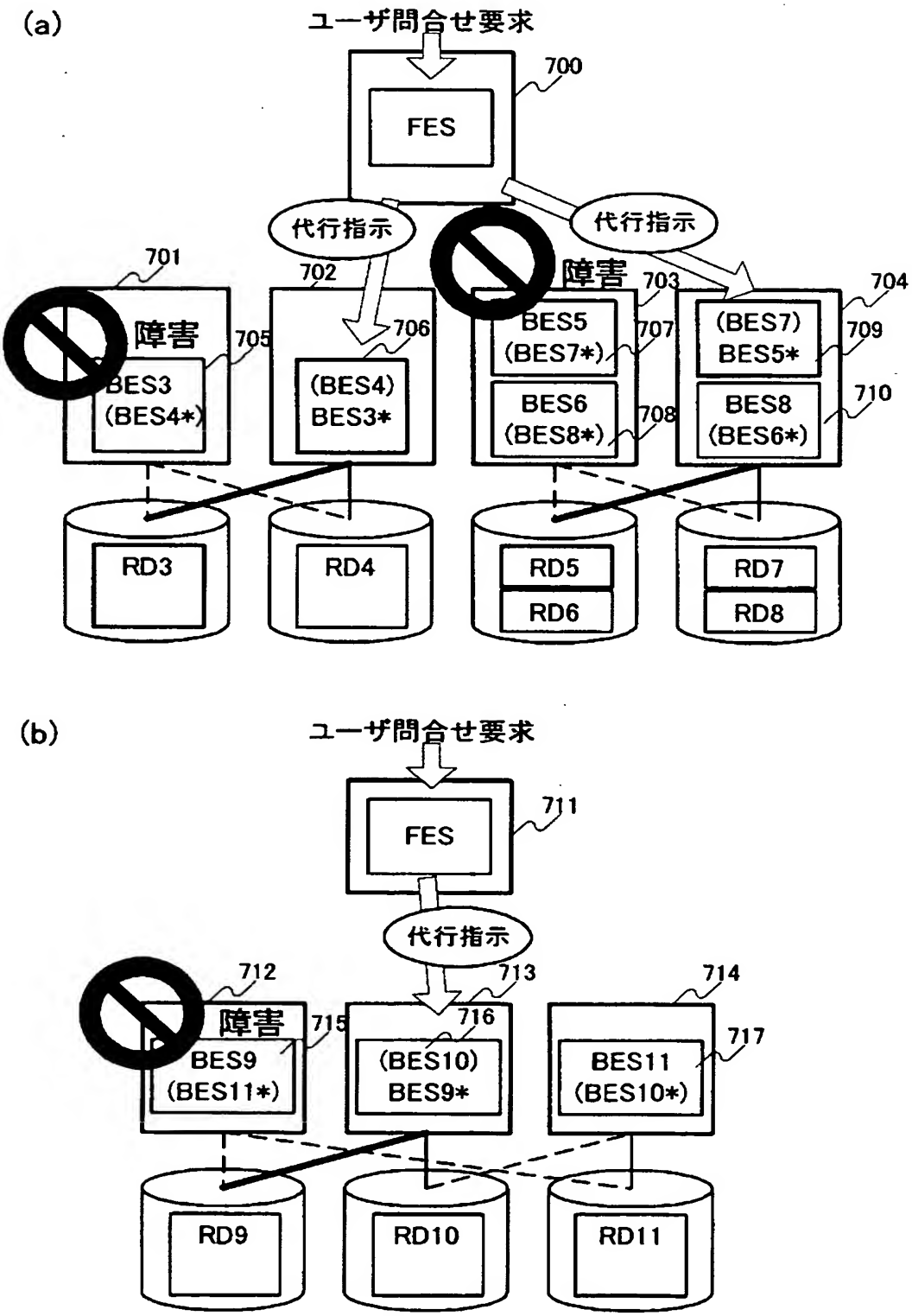
サーバ	代行サーバ	
BES12	BES13	610
BES13	BES12	611
BES14	BES12	612

(e)

サーバ	代行サーバ	代行優先順	
BES15	BES16	1	613
BES15	BES17	2	614
BES15	BES18	3	615
‥	‥	‥	

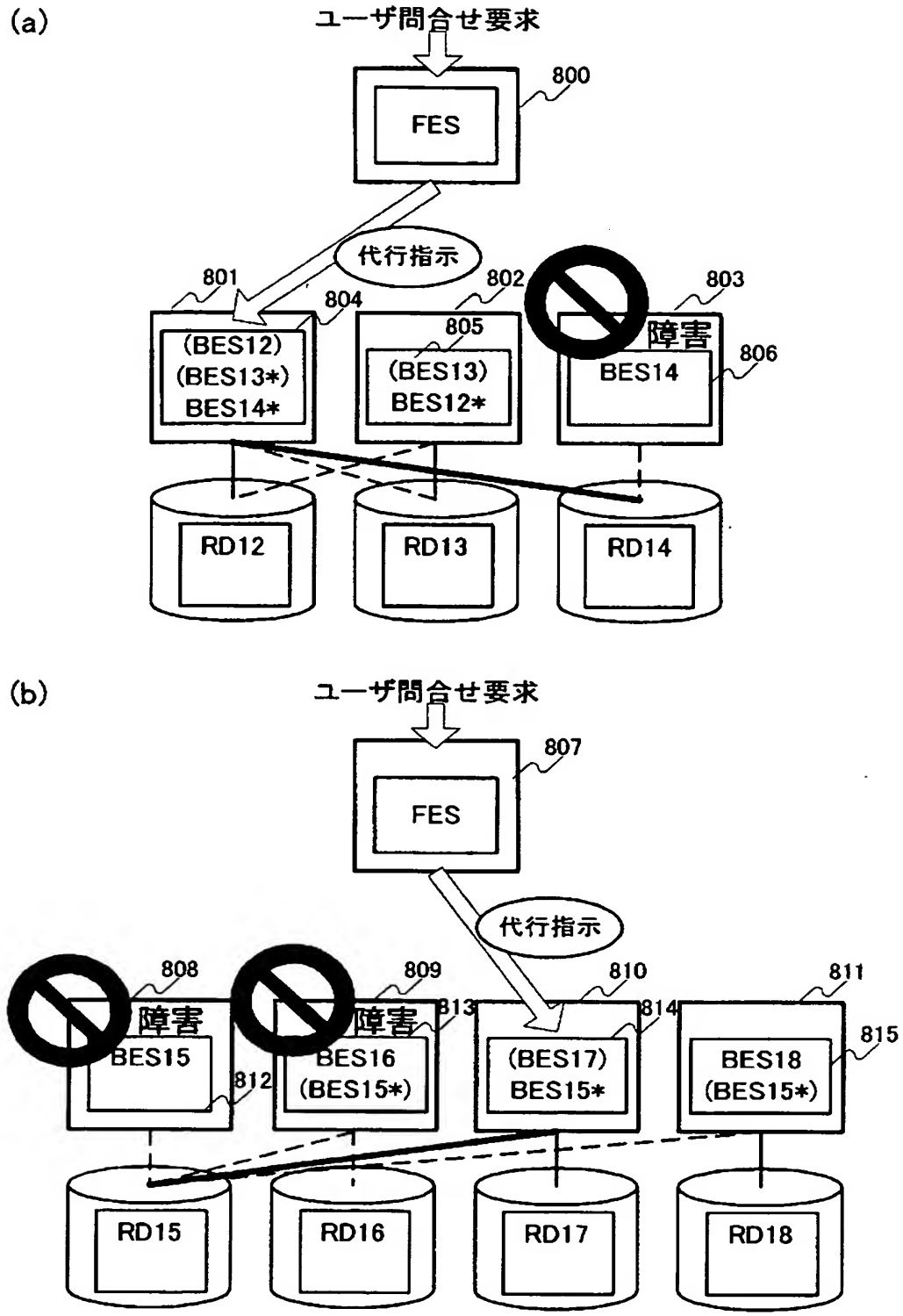
【図７】

【図7】


(a)
ユーザ問合せ要求
700
FES
代行指示
代行指示
障害
701
障害
705
BES3
(BES4*)
702
706
(BES4)
BES3*
703
BES5
(BES7*)
707
BES6
(BES8*)
708
704
(BES7)
BES5*
709
BES8
(BES6*)
710
RD3
RD4
RD5
RD6
RD7
RD8
(b)
ユーザ問合せ要求
711
FES
代行指示
712
障害
715
BES9
(BES11*)
713
716
(BES10)
BES9*
714
BES11
(BES10*)
717
RD9
RD10
RD11

【図８】

【図8】


(a)
ユーザ問合せ要求
FES
800
代行指示
801
802
803
804
805
(BES12)
(BES13*)
BES14*
(BES13)
BES12*
障害
BES14
806
RD12
RD13
RD14
(b)
ユーザ問合せ要求
807
FES
代行指示
808
809
810
811
障害
BES15
障害
BES16
(BES15*)
813
814
(BES17)
BES15*
BES18
(BES15*)
815
812
RD15
RD16
RD17
RD18

【書類名】　要約書

【要約】

【課題】

データベース管理システム、特に、Shared nothingアーキテクチャを用いた並列データベース管理システムにおいて、通常時未稼動状態である待機専用のリソースを必要としない系切り替え機能を提供する。

【解決手段】

ダウン状態の他のＤＢサーバの処理を代行する関係を示す情報を登録しておき、ユーザからの問合せを受け付け、あるＤＢサーバに処理を要求する際に、そのＤＢサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更する。その要求に代行するための指示を付加する。上記要求を受け取ったＤＢサーバは、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの替わりにデータ処理を行う。

これにより、通常時に稼動していない待機専用のリソースを有することなく、障害発生時にＤＢ処理サービスをすぐに再開することが可能なデータベース処理方法およびシステムを提供することができる。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００３－１１５１８５
受付番号	５０３００６５１９６４
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　４月２２日

＜認定情報・付加情報＞

【提出日】　平成15年 4月21日

次頁無

特願２００３－１１５１８５

出　願　人　履　歴　情　報

識別番号　　［０００００５１０８］

１．変更年月日　　１９９０年　８月３１日
［変更理由］　　新規登録
住　所　　東京都千代田区神田駿河台４丁目６番地
氏　名　　株式会社日立製作所